WHEN SHOULD WE ADMIT GENERATED NODES? UTILITY-GUIDED FEW-SHOT LEARNING ON TEXT-ATTRIBUTED GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generated nodes offer a promising way to augment limited supervision in text-attributed graphs, but their usefulness depends on whether they improve learning rather than merely appear plausible. We study this problem as selective synthetic-node admission, where accepting a candidate determines both its use as supervision and its participation in message passing. We introduce CSNA, which ranks candidates using semantic and structural signals learned from validation-derived utility, retains a selected top fraction, and removes rejected nodes together with their attachment edges. Across Cora and PubMed, CSNA outperforms admit-all across all $1/2/5/10$-shot settings, with one-shot accuracy gains of $+6.38$ and $+5.33$ percentage points, respectively. Under 40% synthetic-label corruption, the Cora one-shot gain reaches $+17.80$ points. Together, these results show that the value of synthetic augmentation lies not only in what is generated, but in what is ultimately admitted.

1 INTRODUCTION

Text-attributed graphs combine textual content with relational structure, providing complementary information for node classification. Graph neural networks (GNNs) exploit relations between nodes (Kipf and Welling, 2017; Veličković et al., 2018), while language models provide richer node representations (Chien et al., 2022; Zhao et al., 2023; He et al., 2024). In few-shot settings, however, learning a decision boundary remains difficult when only a few labeled nodes are available for training. LLM4NG addresses this scarcity by generating class-conditioned textual nodes and attaching them to the observed graph (Yu et al., 2025). The resulting candidate pool creates a further learning problem: which generated nodes should be admitted as supervision, and allowed to influence the graph?

Answering this question requires distinguishing a candidate's plausibility from its contribution to learning. A fluent description may carry an inappropriate label, repeat information already present in the support set, or connect to unsuitable neighbors. An atypical candidate, meanwhile, may provide useful supervision despite weak similarity to its class prototype. The graph setting makes this distinction consequential: an admitted node influences training through both its supervised loss and the messages it passes to other nodes. Reducing its loss weight leaves its structural influence in place. Admission therefore involves a joint decision about supervision and graph participation.

We introduce CSNA (selective synthetic-node admission), which separates generation from the decision to put a candidate in the graph. A ranking score uses semantic and structural cues and a leave-one-in validation-utility target; the principal policy keeps a validation-selected top fraction and removes rejected nodes and their attachment edges before message passing. The interface does not require generator internals. Soft weighting is a separate variant: it changes loss coefficients while leaving every candidate in the graph.

We study CSNA in a *validation-assisted few-shot* setting: the student trains on a small labeled support set, while disjoint validation data guide utility estimation, policy selection, and early stopping. Candidate features draw on support labels alone, so the ranking never observes the labels that later choose the policy. This protocol also delimits the claim we make. Admission earns its cost when the real support is scarce or the synthetic labels are unreliable, and it earns much less once a larger support set already fixes the decision boundary.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Cora and PubMed carry the core text-generation evidence, and there CSNA improves mean accuracy over admit-all in all eight $1/2/5/10$-shot cells. The advantage concentrates where supervision is thinnest: the paired one-shot differences are $+6.38$ on Cora and $+5.33$ on PubMed, whereas by ten shots the PubMed interval already includes zero. Corrupting the synthetic labels widens the gap instead of closing it, with the Cora one-shot difference reaching $+17.80$ at a 40% flip rate and $+7.21$ when the same rate arrives through a mixed pool of realistic generation failures. The two remaining datasets play narrower roles. Citeseer extends the admission rule to feature prototypes and produces the largest single gain we report ($+18.23$ at $K=1$), but its candidates are prototypes rather than generated text, so it is not textual LLM evidence. ogbn-arxiv checks behavior at scale, where accuracy rises while macro-F1 falls at $K=1$ and $K=2$, because one global keep threshold over-admits frequent classes. The picture is not one-sided elsewhere either: random retention and GCE overtake CSNA in some higher-K cells, and a matched leave-one ablation that reproduces the Cora $K=1$ mean shows that no individual cue is required in every setting.

Our contributions are threefold:

- We treat synthetic-node admission as a graph decision: a candidate contributes both supervised loss and message passing, or neither. Loss weighting is a different intervention.
- We propose CSNA, which ranks candidates for hard admission from validation-derived utility, not as a new generator and not as a calibration method.
- We evaluate when that decision helps, namely under scarce real support and unreliable synthetic labels, using Cora/PubMed as core text evidence, Citeseer as a feature-level extension, and ogbn-arxiv as a scale check.

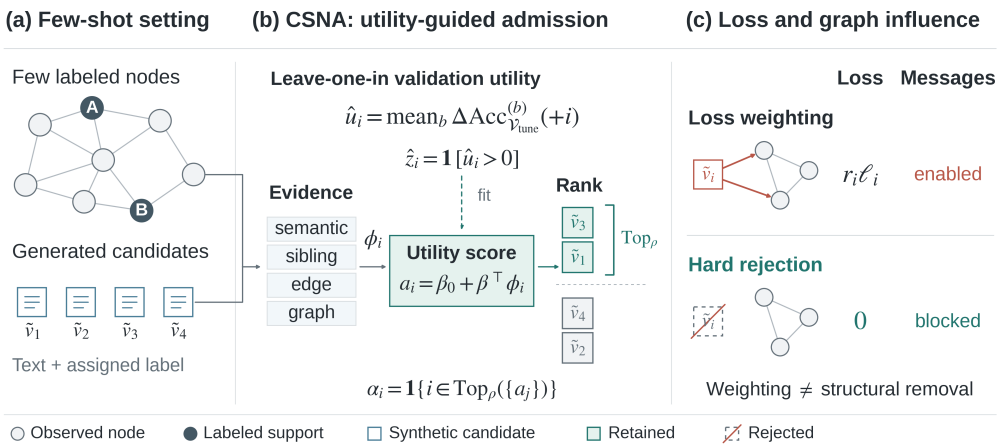

Figure 1: Selective admission versus loss weighting. (a) Generated candidates supplement a text-attributed graph with limited labeled support. (b) CSNA ranks candidates and retains a validation-selected top fraction. (c) Reweighting a node's loss leaves its messages in the graph; hard rejection removes both the supervised term and the attachment edges. The ranking panel is schematic.

2 Related Work

Representation learning on text-attributed graphs. Work on text-attributed graphs treats text and structure as complementary views of the same object. TextGCN converts a corpus into a document–word graph and classifies by propagating over word co-occurrence (Yao et al., 2019). Later methods keep a pretrained language model in the loop: GIANT learns features through self-supervised neighborhood prediction (Chien et al., 2022), GraphFormers nests GNN layers inside a Transformer so that text and neighborhood are encoded jointly (Yang et al., 2021), and GLEM alternates a language model with a GNN under a variational objective that scales to large graphs (Zhao et al., 2023). TAPE instead prompts an LLM for explanations that are interpreted back into node features (He et al., 2024), while broader studies ask where LLMs genuinely help on graph tasks (Chen et al., 2023). Sentence encoders such as Sentence-BERT remain the standard map from node text to fixed-dimensional vectors (Reimers and Gurevych, 2019), and we adopt the same interface: CSNA sees candidates only

as embeddings with assigned labels. All of this improves the representation of nodes that already exist; our question starts once new nodes are proposed and something must decide whether they belong.

Generating new nodes as supervision. Augmenting a graph with synthesized nodes predates LLMs. GraphSMOTE interpolates minority-class nodes in embedding space and trains an edge generator to connect them, making explicit that a synthetic node needs both features and a place in the topology (Zhao et al., 2021). LLM4NG replaces embedding arithmetic with class-conditioned text generation: an LLM writes exemplar documents per label, an edge predictor attaches them, and the augmented graph supplies few-shot supervision (Yu et al., 2025). That is the pipeline we build on, and since its edge predictor already filters some attachments, we do not claim that prior work admits candidates indiscriminately. Two recent systems push further on reliability: SaVe-TAG interpolates at the level of text and adds confidence-based edge assignment for long-tailed graphs (Wang et al., 2026), while GOE-LLM identifies or generates out-of-distribution nodes to expose a classifier to the open world (Xu et al., 2025). Both intervene at the point of attachment. CSNA differs in what the decision ranges over and in what supervises it: we estimate per-node utility from a downstream learning outcome and use it to decide membership in the graph, rather than refining connectivity for candidates retained by default. Reading only embeddings and labels, the score needs no generator internals.

Few-shot and semi-supervised learning on graphs. When labels are scarce, a common response is to manufacture supervision from the unlabeled graph. M3S uses multi-stage self-training with clustering to check its pseudo-labels (Sun et al., 2020), Meta-PN meta-learns a label-propagation strategy (Ding et al., 2022), and CGPN propagates through a contrastive Poisson network for extremely limited supervision (Wan et al., 2021). A parallel line drops labels during pretraining: DGI maximizes mutual information between local patches and a global summary (Veličković et al., 2019), MVGRL contrasts structural views (Hassani and Khasahmadi, 2020), and GRACE contrasts corrupted copies of the graph (Zhu et al., 2020). These methods share our regime but not our object of study, since the supervision they add attaches to existing nodes. Admitting a generated node changes the node and edge sets, altering the propagation operator rather than only the labels flowing through it. Our portability study uses backbones from this literature: GCN, GAT, GraphSAGE, and APPNP (Kipf and Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017; Gasteiger et al., 2019).

Learning from unreliable labels. Since generated labels are not guaranteed correct, the noisy-label literature is our closest methodological neighbor. Generalized cross entropy bounds the influence of any single wrong label (Zhang and Sabuncu, 2018), co-teaching trains two networks that select small-loss examples for each other (Han et al., 2018), and reweighting methods learn per-example coefficients from a clean validation signal, by meta-optimization (Ren et al., 2018) or through a learned weighting function (Shu et al., 2019). NoisyGL benchmarks this family on graphs and finds that structure can spread an error rather than contain it (Wang et al., 2024). The CANE preprint is closest in spirit to our reliability estimate: LLM annotation error is region-dependent, not merely class-dependent, and cluster-conditional reliability can be estimated without ground-truth labels (Thapaliya et al., 2026). CANE annotates existing nodes label-free, whereas we score nodes that do not yet exist and pay with validation labels. More consequentially, every method here adjusts an objective: a down-weighted node still occupies the graph and still sends messages, which is exactly the influence hard rejection removes. We therefore treat robust losses as comparators and complements rather than as alternatives that subsume admission.

Calibration, selective prediction, and risk control. Our score could be read as a probability, so it is worth stating which guarantees we inherit. Temperature scaling and related post-hoc maps correct systematic over-confidence (Guo et al., 2017), proper scoring rules characterize what such probabilities should optimize (Gneiting and Raftery, 2007), and verified calibration analyzes when a measured calibration error can be trusted (Kumar et al., 2019). Selective prediction adds a reject option and studies the risk–coverage trade-off (El-Yaniv and Wiener, 2010), and conformal risk control gives distribution-free guarantees under a valid calibration procedure (Angelopoulos et al., 2024). Synthetic data has itself been used to improve calibration, with error bounds for generated text that fills gaps in validation coverage (Ba et al., 2024). Our position is deliberately modest. CSNA's principal policy keeps a validation-selected top fraction, which depends only on score ordering;

positive-temperature scaling is strictly increasing and cannot change the retained set, so the gains are not a calibration effect. A probability interpretation matters only for soft weighting, and we claim a risk certificate for neither.

Positioning. Prior work thus supplies better representations, generators, attachment rules, and losses for unreliable labels, each implying that not all generated content deserves equal treatment. What is missing is an explicit, learned answer to the admission question itself: should a given candidate enter the training graph at all? CSNA answers it by estimating utility from the downstream learner and enforcing the decision structurally, keeping the choice auditable per node and separable from both the generator and the student.

3 PRELIMINARIES

3.1 GRAPHS AND SUPERVISION

Let $G = (V, E, X)$ be a text-attributed graph with node embeddings $X \in \mathbb{R}^{|V|\times d}$ and labels in $\{1, \dots, C\}$. The labeled support set $S \subset V$ contains $K \geq 1$ nodes per class. We write $S_c = \{v \in S : y_v = c\}$, so that $|S_c| = K$ and $|S| = CK$. Observed adjacency and node features are available transductively. Additional validation labels are partitioned into a tuning set V_{tune} and a disjoint selection set V_{sel}, both separate from the support and test sets. The tuning set supervises utility estimation; the selection set determines the admission policy and student early stopping. Test labels are used only for final evaluation. Thus, *few-shot* refers to the student's labeled real support: the validation labels constitute additional supervision.

3.2 SYNTHETIC AUGMENTATION AND GRAPH PARTICIPATION

A generator supplies a candidate pool $P = \{P_i\}_{i=1}^{N_s}$, where $P_i = (\tilde{x}_i, \tilde{y}_i)$ contains a candidate embedding and its assigned class. For an index set I, write $P[I] = \{P_i : i \in I\}$. Candidate i corresponds to a new node $\tilde{v}_i$. This interface accommodates both generated text embedded by a language model and candidates constructed directly in feature space.

We distinguish a candidate's presence in the graph from its contribution to the supervised loss. Let $m_i \in \{0, 1\}$ indicate structural admission, let $w_i \in [0, 1]$ denote a loss weight, and define $I(m) = \{i : m_i = 1\}$. For an attachment cap κ, the augmented graph $G_m^{(\kappa)}$ contains the original graph and the candidates in $I(m)$, attached using the augmentation pipeline's rule. Rejection removes both the candidate and its incident attachment edges. We write G_m when the attachment configuration is fixed.

Let $p_\theta(y \mid v, G_m)$ denote the graph learner's class probability. The training objective is

$$\begin{aligned}\mathcal{L}(\theta; m, w) = &-\sum_{v\in S} \log p_\theta(y_v \mid v, G_m) \\ &-\sum_{i\in I(m)} w_i \log p_\theta(\tilde{y}_i \mid \tilde{v}_i, G_m).\end{aligned} \tag{1}$$

Admitting all candidates sets $m_i = w_i = 1$. Soft weighting retains every candidate and varies w_i, whereas hard admission varies m_i and sets $w_i = 1$ for retained candidates. Consequently, setting $w_i = 0$ removes a candidate's direct loss contribution while leaving its messages in the graph whenever $m_i = 1$.

4 UTILITY-GUIDED SYNTHETIC-NODE ADMISSION

CSNA learns which generated nodes to admit into the training graph. Its central contribution is to connect candidate selection to an observed learning outcome: whether adding a candidate improves a probe learner's validation accuracy. The selected nodes then contribute both labels and messages to the student. This separates admission from generation while accounting for the two ways synthetic nodes influence graph learning. Algorithm 1 in Appendix C implements this idea in four steps, and the line numbers referenced below point to it: describe candidates, construct utility targets, learn a ranking, and train on the selected augmented graph.

4.1 DESCRIBE CANDIDATES USING SEMANTIC AND STRUCTURAL EVIDENCE

We first describe how each candidate relates to its assigned class and the observed graph (lines 1–4). Both views matter: a candidate can resemble its class in text while having weak support for its proposed connections. We therefore use two semantic signals and two signals about connectivity and neighboring labels.

Let $c_c = |S_c|^{-1} \sum_{v \in S_c} x_v$ be the support prototype of class c, and let $P_c = \{i : \tilde{y}_i = c\}$ contain the indices of candidates assigned to that class. The four signals are

$$q_{\text{sem}}(i) = \cos(\tilde{x}_i, c_{\tilde{y}_i}), \qquad q_{\text{agr}}(i) = \frac{\sum_{j \in P_{\tilde{y}_i} \setminus \{i\}} \cos(\tilde{x}_i, \tilde{x}_j)}{|P_{\tilde{y}_i}| - 1}, \tag{2}$$

$$q_{\text{edge}}(i) = \frac{1}{J_i} \sum_{v \in T_i} p_\psi(i, v), \qquad q_{\text{graph}}(i) = \frac{1}{k_i} \sum_{v \in N^S_{k_i}(i)} \mathbf{1}\{y_v = \tilde{y}_i\}. \tag{3}$$

Each signal supplies a different piece of evidence:

- **Prototype alignment** anchors the candidate to labeled real examples of its assigned class.
- **Candidate agreement** measures consistency with other generated examples of that class, providing additional context when the real support is small.
- **Predicted connectivity** measures how well the candidate matches connection patterns learned from the observed graph.
- **Local label agreement** checks whether nearby support nodes favor the assigned class when all classes can compete.

To obtain the connectivity signal, line 1 trains a two-layer MLP p_ψ on observed real edges and sampled nonedges, without using node-class labels. The set T_i contains the $J_i = \min(20, |S_{\tilde{y}_i}|)$ highest-scoring same-class support nodes. For local agreement, $N^S_{k_i}(i)$ searches all support classes, with written $k = 10$ and $k_i = \min(k, |S|)$. This search allows disagreement with the assigned label to appear in the feature. Among observed node labels, only support labels enter these computations.

Line 4 applies min–max normalization to put the four signals on comparable scales, producing $\phi_i \in \mathbb{R}^4$; Φ stacks these vectors as rows. Their relationship to learning utility is learned in the next steps. Boundary cases are discussed in Section 5 and Appendix C.

4.2 CONSTRUCT SUPERVISION FROM A CANDIDATE'S EFFECT ON LEARNING

The next step supplies a training target for admission (lines 5–13). We ask whether the probe predicts held-out labels more accurately after receiving candidate i. This target connects admission to the downstream task, allowing the score to learn which combinations of evidence are associated with improved predictions.

Lines 5–6 train three support-only probe GCNs, which are reused across candidates. For each candidate, lines 7–11 attach that candidate and train corresponding candidate-added probes. Let $f_S^{(b)}$ and $f_{S+i}^{(b)}$ denote the two probes under initialization b. We compute

$$\begin{aligned} \widehat{u}_i &= \frac{1}{3} \sum_{b=1}^{3} \left[\text{Acc}_{V_{\text{tune}}}(f_{S+i}^{(b)}) - \text{Acc}_{V_{\text{tune}}}(f_S^{(b)}) \right], \\ \widehat{z}_i &= \mathbf{1}\{\widehat{u}_i > 0\}. \end{aligned} \tag{4}$$

The attachment configuration $\mathcal{A}_{\text{probe}}$ is held fixed so that candidates are assessed under the same graph-construction rule. Averaging three initializations is intended to reduce sensitivity to a single training run. The binary target records whether the average change is positive, turning admission-score fitting into a classification problem. Utility targets use V_{tune}; the separate selection set is reserved for choosing the final admission policy and student.

4.3 Learn a ranking from the four signals

Lines 14–17 fit a shared mapping from candidate evidence to the utility target. When the targets contain both classes, an ℓ_2-regularized logistic model produces

$$a_i = \beta_0 + \beta^\top \phi_i. \tag{5}$$

This model lets the contribution of each signal be learned from observed probe outcomes. A linear combination keeps the four-feature scoring rule simple, while regularization discourages large coefficients. If every utility target has the same value, the targets provide no positive-versus-nonpositive contrast; the uniform-feature fallback in line 17 supplies a ranking in this case.

The hard policy uses the logits a_i directly. It therefore needs an ordering of candidates, with no probability threshold to interpret. Temperature scaling preserves this ordering; its role in probability weighting is discussed separately in Section 5.

4.4 Admit candidates and train the student

The final step chooses how much synthetic supervision to retain and how strongly to connect it to the graph (lines 18–26). For each retention fraction ρ, line 20 selects the highest-scoring candidates:

$$m_i^{\mathrm{hard}} = \mathbf{1}\{i \in \mathrm{Top}_\rho(a)\}, \qquad w_i^{\mathrm{hard}} = 1. \tag{6}$$

Here Top_ρ returns retained indices under the run's grouping, count-rounding, and tie-breaking conventions. Retention controls the amount of synthetic supervision. The attachment cap $\kappa = K_{\mathrm{attach}}$ separately controls how many connections each retained node can introduce. We select both because retaining a candidate and determining its connectivity are distinct decisions.

Line 22 builds the graph using only retained candidates. This step removes rejected nodes from both supervision and message passing. Merely setting their loss weights to zero would leave their messages in place. Line 23 then trains the student on the real support and admitted synthetic labels using Equation (1).

We search $\mathcal{R} = \{0.5, 0.7, 0.85\}$ and $\mathcal{K} = \{0, 3, 5\}$. The student is a two-layer GCN with hidden width 32, dropout 0.2, Adam learning rate 0.01, and at most 500 epochs. The selection set V_{sel} supplies early stopping and the accuracy used to compare policies (lines 23–26). This gives policy selection labels that were not used to construct the utility targets or fit the admission score. The principal configuration reports $K_{\mathrm{attach}} = 3$. After selection, the returned student is evaluated on test labels with the policy and checkpoint fixed.

5 Analysis: Ranking, Weighting, and Graph Influence

Two distinctions frame CSNA. First, for $T > 0$ the map $a \mapsto \sigma(a/T)$ is strictly increasing, so temperature scaling cannot change a fixed top-fraction selection (Proposition 3). Hard-admission gains are therefore not a calibration effect. Soft weighting may still use $r_i = \sigma(a_i/T)$ as a loss coefficient while keeping every candidate in the graph.

Second, Equation (1) separates the structural mask m from the loss weights w. Setting $w_i = 0$ removes a candidate's supervised term while leaving its messages in G_m; changing m removes the node and its attachment edges and can alter predictions on other nodes. Appendix B records a fixed-parameter GCN perturbation bound for this route, together with elementary weighting-gap and selective-risk statements that apply only under their stated targets. None of these arguments proves that the measured procedure improves test accuracy. When $|S| \le k$ under the written $k = 10$, balanced support makes $q_{\mathrm{graph}}(i) = 1/C$ for every candidate (Appendix C); low-shot ranking is then not identified with that cue.

6 Experimental Setup

Task and dataset roles. Citation-dataset runs sample $K \in \{1, 2, 5, 10\}$ support nodes per class and then draw 500 validation and $1{,}000$ test nodes. The validation labels are genuine additional supervision, while the candidate features are computed from support labels alone. We therefore

Table 1: **Principal comparison.** Accuracy (%, mean±std, 10 seeds); differences are CSNA minus admit-all. Δ_{F1} is the macro-F1 difference. Roles: Cora/PubMed are core text-generation evidence; Citeseer is a feature-prototype extension; ogbn-arxiv is a scale check (positive-shot track).

Dataset	K	Admit-all	CSNA	Δ_{Acc}	Δ_{F1}
Cora	1	72.27±2.44	**78.65±1.65**	+6.38	+7.50
	2	73.23±3.63	**77.65±1.97**	+4.42	+5.08
	5	77.08±1.09	**79.48±1.71**	+2.40	+2.85
	10	79.27±1.99	**80.49±1.76**	+1.22	+1.47
PubMed	1	69.06±5.08	**74.39±3.73**	+5.33	+5.29
	2	71.55±4.90	**75.64±3.45**	+4.09	+4.02
	5	75.00±2.69	**76.86±2.36**	+1.86	+1.74
	10	77.56±2.30	**77.82±2.04**	+0.26	+0.15
Citeseer	1	29.84±6.66	**48.07±5.59**	+18.23	+17.86
	2	41.13±4.40	**56.67±1.64**	+15.54	+14.17
	5	49.72±4.09	**59.59±2.53**	+9.87	+9.52
	10	57.30±3.87	**61.15±3.92**	+3.85	+2.84
ogbn-arxiv	1	52.04±4.36	**54.50±3.87**	+2.46	-1.28
	2	54.84±2.83	**56.83±2.63**	+1.99	-1.18
	5	55.60±2.34	**58.04±2.26**	+2.44	+0.68
	10	56.76±1.04	**58.80±1.43**	+2.04	+0.84

describe the protocol as *validation-assisted K-shot learning* rather than as a budget of only CK annotations, and we keep that distinction visible in every comparison. The four datasets play three roles. Cora (2,708 nodes, 7 classes) and PubMed (19,717 nodes, 3 classes) supply the **core text-generation evidence**, using the text shipped with LLM4NG. Citeseer (3,327 nodes, 6 classes) is a **feature-level extension**, because its synthetics are support prototypes rather than generated text. ogbn-arxiv (169,343 nodes, 40 classes) serves as a **scale and generalization check**, run on a 10-seed positive-shot track rather than the official OGB leaderboard protocol (Hu et al., 2020). Table 18 varies n_{val} while giving both methods the same extra labels; those 5-seed means diagnose how sensitive the 500-validation headline is, and they do not replace it.

Scope. The main pipeline consists of Table 1 with its paired tests, the Cora/PubMed selector and weighting studies, the matched four-signal ablation, and the utility tables. Everything else is supplementary and descriptive, including the corruption sweep and the unmatched ladder in Appendix F. One number warrants an explicit caution: the unmatched Full accuracy of 71.05 in Appendix G comes from a historical configuration, and we do not use it to explain the headline 78.65. Unless stated otherwise, means are taken over 10 seeds; the backbone, retention, and equal-validation studies use 5 seeds, the in-sample utility diagnostic uses 3, and the held-out utility ranking uses 10. Paired 95% Student-t intervals, Wilcoxon tests, and d_z are reported only for the principal CSNA-versus-admit-all seeds, no cell is starred, and Appendix H records the arithmetic conventions behind the reported differences.

7 Results

7.1 Does admission improve the principal augmentation pipeline?

Table 1 favors CSNA in all eight Cora/PubMed cells, and the size of that advantage tracks how little real supervision the student receives. On Cora the gain falls from $+6.38$ at $K{=}1$ to $+1.22$ at $K{=}10$, and on PubMed it falls from $+5.33$ to $+0.26$. The paired intervals in Table 12 tell the same story with uncertainty attached: at one shot they are $+6.38$ $[4.17, 8.59]$ on Cora and $+5.33$ $[2.41, 8.25]$ on PubMed, both well separated from zero. At ten shots Cora stays positive but small ($+1.22$ $[0.58, 1.86]$), whereas the PubMed interval $+0.26$ $[-0.96, 1.48]$ contains zero, and we therefore do not read that cell as an improvement.

The remaining two datasets answer narrower questions. Citeseer applies the same admission rule to feature prototypes, and its one-shot interval of $+18.23$ $[14.46, 22.00]$ is the largest in the paper; because those candidates are prototypes rather than generated documents, the cell extends the method's reach without adding textual LLM evidence. ogbn-arxiv then tests the rule at scale, and the

Table 2: **Competing selectors: the separate baseline suite.** Entries are the recorded baseline-minus-CSNA accuracy differences (pp; 10 seeds), not contrasts pooled with Table 1. Negative favors CSNA; positive favors the comparator. The historically named oracle is an embedding-utility proxy, not a true oracle.

	Cora				PubMed			
Comparator	K=1	K=2	K=5	K=10	K=1	K=2	K=5	K=10
Co-teaching	-0.36	+0.46	+1.06	+0.54	-3.10	-1.58	+0.38	+0.21
CSNA-soft (isotonic)	-0.53	+0.59	+0.87	+0.34	-2.60	-1.57	-1.05	-0.07
Edge-only	-2.35	-0.59	+0.31	+0.52	-4.31	-3.26	-3.64	-0.51
GCE	-0.36	+0.46	+1.17	+0.54	-3.10	-1.58	+0.38	+0.21
Graph-only	-0.84	-0.88	+0.31	+0.05	-3.11	-3.10	-4.38	-0.84
Admit-all	-6.38	-4.41	-2.40	-1.22	-5.33	-4.17	-1.86	-0.26
Embedding proxy	-5.08	-2.66	-0.70	-0.46	-10.01	-8.49	-6.67	-2.37
Random retention	-0.75	+0.64	+0.69	+0.41	-3.45	-2.61	-1.24	+0.63
Semantic-only	-4.34	-1.94	-0.45	+0.28	-4.93	-2.73	-4.10	-1.06

two metrics disagree. Accuracy gains range from $+1.99$ to $+2.46$, yet macro-F1 decreases by 1.28 and 1.18 points at K=1 and K=2, and all four ogbn F1 intervals include zero (Table 13). The class-wise diagnostic in Table 15 explains the discrepancy: the candidate pool is balanced by construction, but a single global keep threshold admits 82.8% of the frequent tertile against only 32.6% of the rare tertile at K=1, and admission rate correlates with class count (Spearman $\rho = +0.62$). The rare tertile consequently loses 5.01 points of F1 and 12.36 points of recall, while the frequent tertile gains $+2.00$ F1.

Ablating the four signals under the same matched protocol (Tables 16–17) reproduces the Cora K=1 Full accuracy of 78.65 ± 1.65, which confirms that the ablation and the headline share one pipeline. No single cue is load-bearing everywhere. The largest Planetoid drop comes from removing candidate agreement on PubMed (-1.88 at K=1 and -2.56 at K=5), and semantic evidence alone is weaker than the full score on Cora K=1 (77.24) as well as on all four ogbn cells. The utility-only row is a grouped diagnostic that isolates leave-one-in helpfulness, and we do not present it as a headline method.

The low-shot gains cannot be attributed to the graph cue. Under the written support-only rule with k=10, any support set with $|S| \leq k$ forces $q_{\text{graph}}(i) = 1/C$ for every candidate, which is exactly the situation at Cora/PubMed K=1 and PubMed K=2. Whatever improves those cells therefore comes from semantic, agreement, and edge ranking, combined with the structural removal of rejected nodes in Equation (1). The matched ablation agrees: dropping the graph cue on Cora K=1 yields 78.79 against the Full 78.65, a difference inside seed noise, while dropping agreement moves PubMed the most (-1.88 at K=1, -2.56 at K=5). One caution carries over to every one of these comparisons: leave-one-in utility probes a single candidate at a time, so it cannot establish that the jointly admitted set behaves additively.

7.2 How competitive are simpler selectors?

Table 2 collects the Cora/PubMed selector suite, and it draws the boundary of the method's advantage more sharply than the principal table does. At K=1, CSNA exceeds every listed alternative on both datasets, including the simpler selectors that use only one of its cues. That ordering does not survive at higher shots: random retention is ahead by $+0.64$ on Cora K=2 and by $+0.63$ on PubMed K=10, and GCE is ahead by $+1.17$ on Cora K=5. Read together, these cells locate the regime where deliberate admission pays off, namely scarce real support or unreliable synthetic labels, and they show that a cheap retention rule is competitive once neither condition holds.

A separate weighting study points the same way without supporting dominance: validation tuning exceeds uniform weighting in all four of its cells, yet a learned mixer is higher at both K=5 cells (Table 7). Additional corruption, portability, retention-budget, utility-ranking, and robust-loss studies appear unchanged in Appendix A.

8 Discussion and Limitations

Taken together, the experiments place CSNA in a specific regime rather than establishing a generally dominant selector: it helps most when the real support is scarce or the synthetic labels are unreliable, and it fades when neither condition holds. Three costs qualify even that claim. First, admission is not a free filter, because it consumes validation labels and compute. The one-seed CPU audit in Table 10 puts end-to-end CSNA at roughly $5\times$ admit-all on Cora and $3\times$ on PubMed, with student training itself unchanged and the overhead falling entirely on score extraction and policy search. Second, that dependence on validation labels becomes visible when they are scarce: at small n_{val}, CSNA loses several Cora/PubMed cells (Table 18). Third, the supervision behind the score is a proxy. Leave-one-in utility probes candidates one at a time and certifies nothing about the jointly admitted set, and its in-sample AUROC of 0.786 is optimistic next to the held-out 0.675 ± 0.162 in Table 6. The clearest failure mode is structural rather than statistical: on ogbn-arxiv a single global keep threshold under-admits rare classes and macro-F1 falls at $K{=}1$ and $K{=}2$, so a class-balanced keep rule is the natural next step, although we do not adopt it as a headline policy here. Historical unmatched ablations remain available in Appendix G for comparison with the matched runs.

9 Conclusion

Admitting a generated node is a decision about the graph, not only about the loss. The intervention that produces our gains is hard admission rather than loss weighting or temperature scaling: a rejected candidate leaves the objective and the graph together, while a down-weighted one keeps passing messages. Cora and PubMed supply the core text-generation evidence for that claim, Citeseer extends the same admission rule to feature prototypes, and ogbn-arxiv shows at scale how accuracy gains can coexist with a low-shot macro-F1 drop. The value of synthetic augmentation therefore depends on what a pipeline admits, and it depends most when real supervision is scarce or the generated labels are unreliable. The question we leave open is how to preserve rare classes once a single global ranking decides what to keep.

AI Use Statement

AI was used for retrieval and discovery (e.g., finding related work). The LLM-based node generation evaluated in the experiments is part of the method under study and is described in the paper.

Ethics Statement

Synthetic supervision can reproduce generator biases or discard atypical but informative examples, and our admission score estimates a task-dependent utility proxy rather than the truth of generated content. The ogbn-arxiv audit shows that global ranking suppresses rare classes even when the candidate pool is balanced, so class-level effects should be inspected before deployment. We do not validate admission for high-stakes decisions about individuals, and any released graph text, model outputs, or checkpoints must respect their licenses and access conditions.

Reproducibility Statement

Appendix B contains the complete analytical statements, assumptions, and proofs, and Appendix C records label provenance, the full procedure in Algorithm 1, and the recorded hyperparameters. Appendix A and Appendices D–G hold the remaining experimental tables together with their study-specific comparators and qualifications, and Appendix H states the evidence-selection and arithmetic conventions. Paired intervals are computed from seed-level test metrics for the principal comparison. Two limits remain: candidate-level scores, exact split masks, and complete executable configurations for the frozen principal track are unavailable, and we claim neither a public code release nor an independent reproduction.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Yang Ba, Michelle V. Mancenido, and Rong Pan. Fill in the gaps: Model calibration and generalization with synthetic data. In *Proceedings of EMNLP*, pp. 17211–17225, 2024. https://aclanthology.org/2024.emnlp-main.955/.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. Exploring the potential of large language models (LLMs) in learning on graphs. *arXiv preprint arXiv:2307.03393*, 2023. https://arxiv.org/abs/2307.03393.

Eli Chien, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, Jiong Zhang, Olgica Milenkovic, and Inderjit S. Dhillon. Node feature extraction by self-supervised multi-scale neighborhood prediction. In *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=KJggliHbs8.

Kaize Ding, Jianling Wang, James Caverlee, and Huan Liu. Meta propagation networks for graph few-shot semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. https://doi.org/10.1609/aaai.v36i6.20605.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. https://jmlr.org/papers/v11/el-yaniv10a.html.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019. https://arxiv.org/abs/1810.05997.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. https://doi.org/10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 1321–1330, 2017. https://proceedings.mlr.press/v70/guo17a.html.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017. https://papers.neurips.cc/paper/6703-inductive-representation-learning-on-large-graphs.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018. https://arxiv.org/abs/1804.06872.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 4116–4126, 2020. https://proceedings.mlr.press/v119/hassani20a.html.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Harnessing explanations: LLM-to-LM interpreter for enhanced text-attributed graph representation learning. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/1766d75b077b66457040e4661771aec5-Abstract-Conference.html.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, 2020. https://arxiv.org/abs/2005.00687.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. https://openreview.net/forum?id=SJU4ayYgl.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 2019. https://papers.nips.cc/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. https://aclanthology.org/D19-1410/.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 4334–4343, 2018. https://proceedings.mlr.press/v80/ren18a.html.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-Weight-Net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, 2019. https://arxiv.org/abs/1902.07379.

Ke Sun, Zhouchen Lin, and Zhanxing Zhu. Multi-stage self-supervised learning for graph convolutional networks on graphs with few labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. https://arxiv.org/abs/1902.11038.

Safal Thapaliya, Jiatan Huang, and Chuxu Zhang. Where LLM annotators fail: Label-free learning on graphs with LLMs. *arXiv preprint arXiv:2605.27913*, 2026. https://arxiv.org/abs/2605.27913.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. https://openreview.net/forum?id=rJXMpikCZ.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019. https://openreview.net/forum?id=rklz9iAcKQ.

Sheng Wan, Yibing Zhan, Liu Liu, Baosheng Yu, Shirui Pan, and Chen Gong. Contrastive graph Poisson networks: Semi-supervised learning with extremely limited labels. In *Advances in Neural Information Processing Systems*, 2021. https://papers.nips.cc/paper/2021/hash/31c0b36aef265d9221af80872ceb62f9-Abstract.html.

Leyao Wang, Yu Wang, Bo Ni, Yuying Zhao, Hanyu Wang, Yao Ma, and Tyler Derr. SaVe-TAG: LLM-based interpolation for long-tailed text-attributed graphs. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. https://doi.org/10.1145/3770854.3780311.

Zhonghao Wang, Danyu Sun, Sheng Zhou, Haobo Wang, Jiapei Fan, Longtao Huang, and Jiajun Bu. NoisyGL: A comprehensive benchmark for graph neural networks under label noise. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. https://papers.nips.cc/paper/2024/hash/436ffa18e7e17be336fd884f8ebb5748-Abstract-Datasets_and_Benchmarks_Track.html.

Haoyan Xu, Zhengtao Yao, Ziyi Wang, Zhan Cheng, Xiyang Hu, Mengyuan Li, and Yue Zhao. Graph synthetic out-of-distribution exposure with large language models. *arXiv preprint arXiv:2504.21198*, 2025. https://arxiv.org/abs/2504.21198.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. GraphFormers: GNN-nested Transformers for representation learning on textual graph. In *Advances in Neural Information Processing Systems*, 2021. https://papers.neurips.cc/paper/2021/hash/f18a6d1cde4b205199de8729a6637b42-Abstract.html.

Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. https://arxiv.org/abs/1809.05679.

Jianxiang Yu, Yuxiang Ren, Chenghua Gong, Jiaqi Tan, Xiang Li, and Xuecang Zhang. Leveraging large language models for node generation in few-shot learning on text-attributed graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):13087–13095, 2025. https://doi.org/10.1609/aaai.v39i12.33428.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018. https://arxiv.org/abs/1805.07836.

Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. Learning on large-scale text-attributed graphs via variational inference. In *International Conference on Learning Representations*, 2023. https://openreview.net/forum?id=q0nmYciuuZN.

Tianxiang Zhao, Xiang Zhang, and Suhang Wang. GraphSMOTE: Imbalanced node classification on graphs with graph neural networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2021. https://doi.org/10.1145/3437963.3441720.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020. https://arxiv.org/abs/2006.04131.

A Extended Empirical Results

The following subsections are moved from the main results without rewriting.

A.1 What happens when synthetic labels are corrupted?

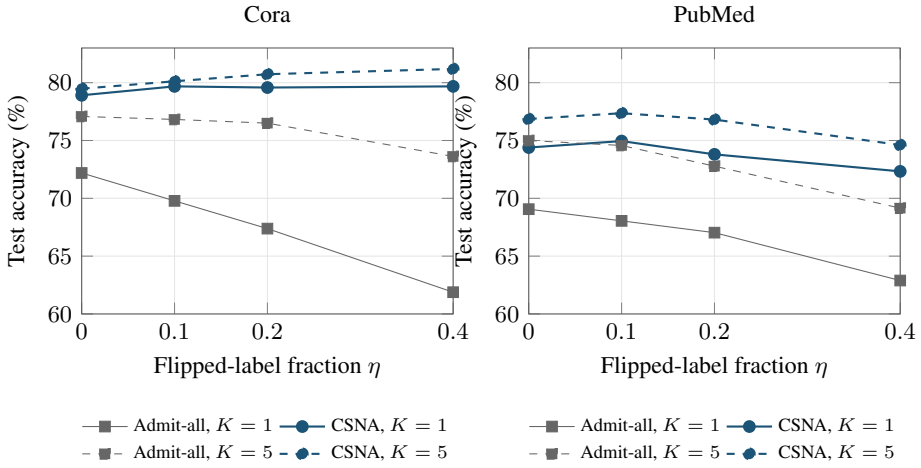

Figure 2: **All corruption settings.** Curves reproduce the 10-seed mean accuracies at $K = 1$ (solid) and $K = 5$ (dashed), for both datasets and all four flip rates. Error bars are not available. Differences across corruption levels do not establish frozen-calibrator robustness.

We evaluate incorrect-label flips at $\eta \in \{0.0, 0.1, 0.2, 0.4\}$ on Cora/PubMed with $K \in \{1, 5\}$, using the same corrupted pool for both methods. CSNA has higher mean accuracy in all 16 cells (Figure 2; full means in Table 22). At Cora $K = 1$, admit-all changes from 72.19 to 61.88 between zero and 40% corruption, whereas CSNA changes from 78.91 to 79.68. The endpoint difference is $+17.80$ points. At Cora $K = 5$, it is $+7.57$.

PubMed is not flat: at $K = 1$, CSNA decreases from 74.39 to 72.33, but remains above admit-all's 62.89. Its advantage is also not strictly monotone: $+6.90$ at $\eta = 0.1$ becomes $+6.77$ at $\eta = 0.2$. The sweep supports resilience under this label-flip evaluation. Because the record does not establish whether the reliability fit was frozen across corruption levels, this is not evidence of zero-adaptation transfer from a clean-pool calibrator. We retain each sweep's zero-noise row rather than replacing it with principal-table values.

Random wrong-class flips are an artificial proxy for generation errors. We therefore add one mixed failure pool that corrupts 40% of the same shipped LLM4NG synthetics with equal shares of plausible wrong-class labels, class-boundary feature blends, forced wrong-class attachment, near-duplicate support copies, and large feature noise. Both methods see the identical mutated pool (Table 23). CSNA has higher mean accuracy in all four Cora/PubMed $\times\{1, 5\}$ cells. At Cora $K = 1$, admit-all changes from 72.27 to 70.53 and CSNA from 78.65 to 77.74; the endpoint difference is $+7.21$ points, smaller than the $+17.80$ label-flip gap because uniform flips degrade admit-all more severely. Clean rows are those of the principal reproducibility track, not Table 22. The mutations act on a fixed pool; the study does not re-generate text from adversarial prompts.

A.2 Does the admission interface transfer across model choices?

Table 3: **Portability summaries from existing grids.** Accuracy differences are CSNA minus the within-study admit-all comparator (pp). Means are unweighted averages of the available cell differences, not a pooled performance estimate. Backbone: Cora/PubMed $\times\{1, 2, 5, 10\}$, 5 seeds. Encoders and generators: Cora/PubMed $\times\{1, 5\}$, 10 seeds. Full grids appear in Appendix E.

Axis	Choice	Positive cells	Mean Δ	Min Δ	Max Δ
Backbone	GCN	8/8	+4.23	+0.40	+8.02
	GAT	7/8	+4.66	-0.32	+8.56
	GraphSAGE	8/8	+7.21	+1.78	+13.72
	APPNP	7/8	+2.27	-2.38	+6.22
Encoder	MiniLM	4/4	+4.20	+1.86	+7.42
	MPNet	4/4	+5.74	+2.34	+10.83
	E5-base	4/4	+3.72	+1.32	+8.02
	BGE-base	4/4	+3.71	+1.97	+6.02
Generator	LLM4NG	4/4	+3.99	+1.86	+6.38
	Template	4/4	+3.39	+1.52	+6.34
	Prototype	4/4	+4.37	+1.98	+10.56

Table 3 summarizes the existing extension grids, not new training runs. With GCN, GAT, GraphSAGE, and APPNP (Kipf and Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017; Gasteiger et al., 2019), accuracy differences are positive in $30/32$ cells. The exceptions are PubMed $K = 10$ with APPNP (-2.38) and GAT (-0.32); GraphSAGE has the largest single difference ($+13.72$ on Cora $K = 1$). These outcomes establish neither backbone-invariant utility nor a causal explanation of GraphSAGE's sensitivity.

Re-embedding the same text with MiniLM, MPNet, E5-base, and BGE-base gives positive differences in all 16 encoder cells. Replacing the synthetic pool with LLM4NG, template, or support-prototype candidates gives positive differences in all 12 generator cells. These are separate applications of the admission procedure, not demonstrated transfer of one frozen score model. Encoder and generator shifts test the admission procedure on a new pool, not transfer of one frozen score model.

A.3 How much retention is necessary, and what do scores predict?

Table 4: **Cora retention budgets.** Test accuracy (%, 5-seed means), by requested synthetics per class M. Admit-all remains a full-pool reference, so repeated reference values do not mean equal retained-node budgets.

K	Method	$M = 1$	2	4	8	16	32
1	Admit-all	72.66	72.66	72.66	72.66	72.66	72.66
1	CSNA	68.28	71.92	72.78	72.32	72.66	72.66
5	Admit-all	77.52	77.52	77.52	77.52	77.52	77.52
5	CSNA	78.32	77.86	77.94	77.80	77.52	77.52

The budget study (Table 4) shows a retention trade-off. At Cora $K = 1$, retaining only $M = 1$ synthetic per class is worse than the full-pool reference, while $M = 4$ is slightly higher. At $K = 5$, several smaller subsets are competitive. The reference is held at its *full pool* for every displayed M; this is not a matched-retention comparison or a measurement of end-to-end compute savings. The table does not support selecting M from test accuracy.

The equal-validation sweep (Table 18) varies the labeled val prefix rather than retention M. Both methods use the same n_{val} extra labels. CSNA is worse than admit-all at n_{val}=10 in every reported cell, remains below admit-all at n_{val}=25 except on Cora K=1, and is mixed at n_{val}=50 (ahead on both K=1 cells; behind on both K=5 cells). We do not promote a small-val CSNA variant, and we do not replace Table 1 or Table 4 with these means.

Table 5: **In-sample score diagnostic.** Cora $K = 1$, 3 seeds, tune-half fit and evaluation on the same synthetics. Ranking evidence is Table 6.

Target	AUROC	AUPRC	ECE	Brier	Spearman ρ
Embedding proxy	0.415	0.509	0.355	0.370	-0.148
Centroid label match	0.377	0.979	0.294	0.148	-0.051
Structural proxy	0.101	0.310	0.504	0.454	-0.687
Marginal utility	0.786	0.863	0.083	0.170	0.447

Table 5 is in-sample: AUROC 0.786 on Cora $K{=}1$. The ranking claim is the held-out AUROC 0.675 ± 0.162 (Table 6). The in-sample figure was optimistic. Structural compatibility is negatively associated with the same in-sample score.

Table 6: **Held-out utility ranking.** Cora $K{=}1$, 10 seeds. Fit on utility-train, temperature on utility-val, metrics on utility-test. In-sample 0.786 remains in Table 5.

Target	AUROC	AUPRC	ECE	Brier	Spearman ρ
Embedding proxy	0.256±0.137	0.434±0.105	0.309±0.102	0.310±0.060	-0.417±0.233
Centroid label match	0.389±0.186	0.973±0.048	0.134±0.216	0.251±0.017	-0.082±0.139
Structural proxy	0.329±0.343	0.518±0.235	0.320±0.114	0.298±0.054	-0.292±0.590
Marginal utility	0.675±0.162	0.690±0.213	0.132±0.071	0.223±0.036	0.297±0.273

A different audit uses embedding proxies on clean pools and known flip status under corruption (Table 14). It reports clean AUROC above 0.97, but that number is not the utility AUROC. Its scores, targets, and evaluation indices are not sufficiently aligned with Table 5 to combine them into one calibration claim. We therefore conclude that the score is associated with its chosen utility proxy, not that it certifies correctness or that calibration explains the downstream gains.

A.4 How do weighting strategies compare?

Table 7: **Weighting strategies within their own study.** Test accuracy (%, mean±std). The study does not record its seed count; it is not pooled with the principal comparison. Learned weighting is higher in the two $K = 5$ cells.

Method	Cora $K = 1$	Cora $K = 5$	PubMed $K = 1$	PubMed $K = 5$
Admit-all	71.51±2.75	77.57±1.12	66.84±4.57	74.13±2.36
Uniform weights	74.74±1.83	79.02±0.99	70.99±2.85	73.60±2.44
Learned mixer	77.91±1.28	80.54±1.46	70.84±3.94	76.22±1.77
Validation-tuned CSNA	78.19±1.68	80.35±1.75	72.84±3.81	75.93±1.76

This study varies how admitted candidates are weighted rather than whether they are admitted, and it is informative in the same qualified way as the selector suite. Validation tuning exceeds uniform weighting in all four cells, which indicates that the selection labels contribute beyond a fixed weighting scheme, yet a learned mixer is higher in both $K = 5$ cells. Because the study does not record its seed count, we report it separately and do not pool it with the principal comparison in Table 1.

A.5 Does admission complement robust losses?

Table 8: **Admission and robust-loss combinations.** Test accuracy (%, 10-seed means) within the separate combination study. CoT denotes co-teaching. Equal rounded means are not evidence that the loss functions are mechanistically equivalent.

Dataset	K	Admit-all	+GCE	+CoT	CSNA	CSNA+GCE	CSNA+CoT
Cora	1	72.27	72.14	70.77	78.65	78.60	78.60
Cora	2	73.23	72.58	71.87	77.65	77.65	77.65
Cora	5	77.08	77.90	76.67	79.48	79.48	79.48
Cora	10	79.27	79.65	79.38	80.49	80.49	80.49
PubMed	1	69.06	68.44	69.25	74.39	74.39	74.39
PubMed	2	71.55	70.88	70.39	75.64	75.64	75.64
PubMed	5	75.00	73.81	73.22	76.86	76.86	76.86
PubMed	10	77.56	76.98	77.28	77.82	78.46	78.46

In the separate robust-loss combination study, CSNA+GCE and CSNA+co-teaching frequently reproduce the same rounded mean as CSNA; Cora $K = 1$ is slightly lower and PubMed $K = 10$ is higher (Table 8). We therefore make neither an established-complementarity claim nor a claim that robust losses consistently improve admit-all.

B Complete Analytical Statements and Proofs

The analysis separates score error, calibration, selection, and graph perturbation. No proposition is a guarantee that the measured hard-admission procedure improves test accuracy. In particular, an audit target must not change meaning between a theorem and the metric used to discuss it.

B.1 Fixed-predictor weighting error

Consider fixed predictions on a fixed graph, bounded losses $0 \le \ell_i \le B$, weights $r_i \in [0, 1]$, and a specified binary target z_i. Define $R_r = N_s^{-1} \sum_i r_i \ell_i$ and $R_z = N_s^{-1} \sum_i z_i \ell_i$. Then

$$|R_r - R_z| \le \frac{B}{N_s} \sum_i |r_i - z_i| \le B \sqrt{\frac{1}{N_s} \sum_i (r_i - z_i)^2}. \tag{7}$$

The first term is absolute binary prediction error; the second is the Brier score for the same target and sample. Neither is ECE. The bound compares loss weightings of fixed predictions, not separately retrained GNNs, and requires a bounded analytical loss.

Proposition 1 (Bounded-loss weighting gap)**.** *For fixed losses $\ell_i \in [0, B]$, weights $r_i \in [0, 1]$, and targets $z_i \in \{0, 1\}$, define $R_r = N_s^{-1} \sum_i r_i \ell_i$ and $R_z = N_s^{-1} \sum_i z_i \ell_i$. Equation* (7) *holds.*

Proof. By the triangle inequality and the bound on the loss,

$$|R_r - R_z| = \frac{1}{N_s} \left| \sum_i (r_i - z_i) \ell_i \right| \le \frac{1}{N_s} \sum_i |r_i - z_i| \ell_i \le \frac{B}{N_s} \sum_i |r_i - z_i|.$$

Cauchy–Schwarz gives $(N_s^{-1} \sum_i |r_i - z_i|)^2 \le N_s^{-1} \sum_i (r_i - z_i)^2$, proving the second inequality. □

The denominator remains N_s in both risks; R_z is not the conditional risk normalized by the number of positive targets. The proof fixes predictions and the propagation graph. It does not compare independently optimized models or automatically apply to unbounded log loss. A mathematically bounded surrogate, such as $\min\{\ell_i, B\}$, makes the premise explicit but is not asserted to have been used in training.

Absolute prediction error is not ECE. Let $Z \sim \text{Bernoulli}(1/2)$ and $R = 1/2$ almost surely. Then $\mathbb{E}[Z \mid R] = R$, so population score calibration error is zero. Nevertheless,

$$\mathbb{E}|R - Z| = \tfrac{1}{2}, \qquad \mathbb{E}(R - Z)^2 = \tfrac{1}{4}.$$

Thus perfect calibration need not eliminate absolute error or Brier score: uncertainty in the target remains. A binned ECE estimate also depends on the binning and is not interchangeable with either quantity (Guo et al., 2017; Kumar et al., 2019).

B.2 A POPULATION ACCEPTANCE BOUND

Let $R \in [0, 1]$ be a score, $Z \in \{0, 1\}$ its specified target, and $\mu(R) = \mathbb{E}[Z \mid R]$. For an acceptance event $A = \{R \geq \tau\}$ with $a = \mathbb{P}(A) > 0$, define $\delta = \mathbb{E}|\mu(R) - R|$. Then

$$\mathbb{E}[1 - Z \mid A] \leq 1 - \tau + \delta/a. \tag{8}$$

This statement concerns the chosen target. A small binned ECE does not identify δ, and a validation-selected top fraction is not a certified threshold rule.

Proposition 2 (Target-specific selective risk)**.** *Let $R \in [0, 1]$, $Z \in \{0, 1\}$, $m(R) = \mathbb{E}[Z \mid R]$, and $A = \{R \geq \tau\}$ with $a = \mathbb{P}(A) > 0$. Writing $\delta = \mathbb{E}|m(R) - R|$, Equation* (8) *holds.*

Proof. Since A is measurable with respect to R, iterated expectation yields

$$\begin{aligned}
\mathbb{E}[1 - Z \mid A] &= 1 - \mathbb{E}[m(R) \mid A] \\
&\leq 1 - \mathbb{E}[R \mid A] + \mathbb{E}[|m(R) - R| \mid A] \\
&\leq 1 - \tau + \frac{\mathbb{E}[|m(R) - R|\mathbf{1}_A]}{a} \leq 1 - \tau + \frac{\delta}{a}.
\end{aligned}$$

□

The bound becomes $1 - \tau$ under perfect population calibration, for that target. A utility proxy, a label-flip indicator, and a verified class label are different targets. A data-selected retention fraction supplies neither the unknown δ nor a finite-sample upper bound on it. Conformal risk control requires its own valid calibration procedure and assumptions (Angelopoulos et al., 2024); it is not implemented here.

B.3 WHY TEMPERATURE DOES NOT CHANGE FIXED-BUDGET ADMISSION

Proposition 3 (Rank invariance)**.** *For fixed logits, any positive scalar temperature preserves a fixed top-fraction or top-count selection if the candidate pool, grouping, and tie-breaking rule remain fixed.*

Proof. The derivative of $g_T(a) = \sigma(a/T)$ is $g_T(a)(1 - g_T(a))/T > 0$ for finite a and $T > 0$. Therefore $a_i > a_j$ if and only if $g_T(a_i) > g_T(a_j)$, and equality is preserved. Every ranking, including one within each fixed class group, is unchanged. The retained set is consequently unchanged under the stated selection rule. □

Soft weights do change with temperature. A fixed numerical threshold in probability space can also change the selected set. Neither exception establishes a benefit from temperature for the principal rank-based policy. Numerical saturation or a changed tie-breaking implementation is outside the exact mathematical statement and is not an experimentally established mechanism.

B.4 STRUCTURAL REMOVAL IS A DIFFERENT INTERVENTION FROM LOSS WEIGHTING

For a common node universe, write a message-passing layer as

$$H_A^{(\ell+1)} = \sigma_\ell(AH_A^{(\ell)}W_\ell), \qquad H_A^{(0)} = X.$$

An absent candidate can be represented in the common universe with zero connections to retained nodes. A real-node prediction is then compared by restriction to real-node rows. The operator A includes the specified normalization and self-loop convention.

Proposition 4 (Fixed-parameter propagation perturbation)**.** *Suppose* $\|A\|_2, \|A'\|_2 \le 1$, $\|A - A'\|_2 \le \epsilon$*, every* σ_ℓ *is* 1*-Lipschitz in Frobenius norm with* $\sigma_\ell(0) = 0$*, and both networks share* X *and all weights. Then an* L*-layer network satisfies*

$$\|H_A^{(L)} - H_{A'}^{(L)}\|_F \le L\epsilon\|X\|_F \prod_{\ell=0}^{L-1} \|W_\ell\|_2. \tag{9}$$

Proof. Let $D_\ell = \|H_A^{(\ell)} - H_{A'}^{(\ell)}\|_F$ and $w_\ell = \|W_\ell\|_2$. The hypotheses give $\|H_{A'}^{(\ell)}\|_F \le \|X\|_F \prod_{j<\ell} w_j$. Adding and subtracting $AH_{A'}^{(\ell)}W_\ell$ and using submultiplicativity gives

$$D_{\ell+1} \le w_\ell D_\ell + \epsilon w_\ell \|H_{A'}^{(\ell)}\|_F \le w_\ell D_\ell + \epsilon\|X\|_F \prod_{j\le\ell} w_j.$$

Induction from $D_0 = 0$ proves the result. Restriction to real-node rows cannot increase the Frobenius norm. □

This establishes a route by which node admission can affect other representations. It does not bound how much pruning helps, identify harmful nodes, or compare the parameters learned after two separate training runs. More generally, if $U(A)$ denotes the validation effect of a candidate set A, there is no assumption that $U(A) = \sum_{i\in A} U(\{i\})$. The proxy in Equation (4) therefore does not solve a proven additive subset-selection objective.

C Procedure, Supervision, and Implementation Scope

C.1 Which labels enter each component?

Table 9: Label provenance under the declared protocol. Transductive adjacency does not require node-class labels.

Component	Support	Validation	Test
Core reliability features	Support only	None	None
Probe utility targets	Training labels	Tune: evaluate utility	None
Logistic/temperature fitting	Via core features	Tune-derived targets	None
Retention/attachment	Via frozen features	Selection	None
Student early stopping	Training labels	Selection	None
Final metrics	None	None	Reporting

The method uses support-only *features*, not support-only supervision overall. A labeled validation set trains the utility estimator as well as selecting hyperparameters. This additional label use is intrinsic to the reported formulation. It must be counted in comparisons with label-free methods or methods that only use validation for stopping.

Algorithm 1 CSNA hard admission. The tuning set supplies utility targets; the selection set chooses the policy and student checkpoint. The procedure returns the selected student, augmented graph, and policy parameters.

CSNA-HARD$(G, S, P, V_{\text{tune}}, V_{\text{sel}})$

ψ = FIT-EDGE-PREDICTOR(G)
for $i = 1$ **to** N_s
  q_i = CANDIDATE-SIGNALS(i, G, S, P, ψ)
Φ = NORMALIZE$(q_1, \ldots, q_{N_s})$
for $b = 1$ **to** 3
  $f_S^{(b)}$ = TRAIN-PROBE$(G, S, \varnothing; b)$
for $i = 1$ **to** N_s
  G_i = ATTACH$(G, \{P_i\}; \mathcal{A}_{\text{probe}})$
  for $b = 1$ **to** 3
    $f_{S+i}^{(b)}$ = TRAIN-PROBE$(G_i, S, \{P_i\}; b)$
    $d_b = \text{Acc}_{V_{\text{tune}}}(f_{S+i}^{(b)}) - \text{Acc}_{V_{\text{tune}}}(f_S^{(b)})$
  $\widehat{u}_i = \frac{1}{3}\sum_{b=1}^{3} d_b$
  $\widehat{z}_i = \mathbf{1}\{\widehat{u}_i > 0\}$
if $\widehat{z}$ contains both 0 and 1
  (β_0, β) = FIT-LOGISTIC$(\Phi, \widehat{z})$
  $a = \beta_0 \mathbf{1} + \Phi\beta$
else a = UNIFORM-FEATURE-SCORE(Φ)
$\mathcal{H} = \varnothing$
for each $\rho \in \mathcal{R}$
  m = TOP-MASK(a, ρ)
  for each $\kappa \in \mathcal{K}$
    $G_m^{(\kappa)}$ = AUGMENT$(G, P[I(m)]; \kappa)$
    f = TRAIN-STUDENT$(G_m^{(\kappa)}, S, P[I(m)], V_{\text{sel}})$
    $s = \text{Acc}_{V_{\text{sel}}}(f)$
    $\mathcal{H} = \mathcal{H} \cup \{(s, f, G_m^{(\kappa)}, \rho, \kappa)\}$
return SELECT-MAX-ACCURACY$(\mathcal{H})$

Algorithm 2 CSNA under the declared split-validation protocol

Require: Observed graph/features G, X; support S; candidate pool P; disjoint $V_{\text{tune}}, V_{\text{sel}}$.
Train the edge predictor on observed edges/nonedges, without node-class labels.
Form the four core signals from S, P, and the edge predictor.
Normalize the features under the fixed run configuration.
Train support-only and candidate-added probes over 3 initializations.
Compute $\widehat{z}_i$ using only V_{tune}, as in Equation (4).
Fit regularized logistic scores, or use the uniform fallback for degenerate targets.
Fit temperature for the probability/soft-weight variant using tune-derived targets.
Select retention and attachment settings on V_{sel}.
Hard: retain ranked candidates and remove rejected attachments.
Soft alternative: weight synthetic losses without assuming their edges are removed.
Train the student, using V_{sel} for early stopping.
Report test metrics only after all choices are fixed.

C.2 RECORDED CONFIGURATION AND BOUNDARY CASES

The principal configuration is `csna_hard_v21`. It uses the four core signals, a two-layer edge MLP on concatenated embeddings, and a two-layer GCN student with hidden width 32, dropout 0.2, Adam learning rate 0.01, and a 500-epoch cap. The retention grid is $\{0.5, 0.7, 0.85\}$; the attachment grid is $\{0, 3, 5\}$, with 3 reported for the principal configuration. The soft comparator `csna_soft_v2` uses isotonic calibration and is not a clean temperature-only ablation.

For positive-shot semantic features we adopt the support-prototype specification stated in the principal protocol table, rather than merging it with the separate class-description variant. For validation roles we adopt the explicit algorithm's allocation: tune-derived utility targets for reliability fitting, selection-

half labels for retention and stopping. These are the declared method specifications; the aggregate results do not independently verify implementation choices.

The support-only graph signal uses nearest neighbors from *all* support classes. Under the written $k = 10$, if $|S| \le k$ (equivalently $CK \le 10$ for balanced support) the capped neighborhood exhausts support and

$$q_{\text{graph}}(i) = \frac{|S_{\tilde{y}_i}|}{|S|} = \frac{K}{CK} = \frac{1}{C}.$$

This identity then applies to Cora and PubMed at $K = 1$, and PubMed at $K = 2$. No discrimination by this score is possible under that definition. The shipped support-only helper defaults to $k = 5$, so only PubMed $K = 1$ ($|S| = 3$) is guaranteed constant if that default was used; the frozen headline CSV does not independently record which k ran. A within-assigned-class search would instead be identically one. Accordingly, graph-only outcomes from an unmatched ablation cannot validate this signal's discriminative value in those settings. The matched $-$graph column on Cora $K = 1$ is 78.79 versus Full 78.65; any movement there is not identified with a varying graph cue.

Min–max normalization of a constant column, singleton sibling groups, nonzero-vector handling for cosine similarity, rank ties, and global versus class-wise retention require explicit run conventions. The principal hard policy is global top-fraction keep; the ogbn-arxiv class-wise diagnostic in Appendix D records that convention rather than retrofitting it onto unmatched aggregates. Auxiliary centroid-margin, support-density, and perturbation-uncertainty cues were named in the earlier instrument, but are not analyzed as independently validated additions here. The zero-shot reproduction grid also requires a feature construction distinct from support prototypes; it is not evidence for the positive-shot algorithm without that construction.

C.3 Fitting objective, compute, and audit metrics

For the candidate set used in fitting, the stated logistic objective is

$$\min_{\beta_0,\beta} -\sum_i [\widehat{z}_i \log \sigma(a_i) + (1 - \widehat{z}_i) \log(1 - \sigma(a_i))] + \lambda_{\text{LR}} \|\beta\|_2^2.$$

The exact regularization value, fitted temperatures, optimizer details for the edge/probe models, and fitting/evaluation candidate indices are not supplied by aggregate tables. Fitting probabilities and assessing them on the same targets does not establish out-of-sample calibration.

If one candidate-added probe costs C_{probe}, the candidate-dependent utility-fitting cost is $\mathcal{O}(3N_s C_{\text{probe}})$, aside from a reusable support-only baseline when the configuration permits it. Signal extraction, candidate edges, hyperparameter search, and student training add costs. Table 10 times the principal hard path rather than that probe loop: `csna_hard_v21` pays validation weight/hparam search, not leave-one-in probes, each seed.

Table 10: Wall-clock cost of principal hard admission versus admit-all (CPU, seed 0, support-only Planetoid protocol). Admission is score extraction plus validation weight/hparam search. Student time is the final GCN. Peak is Linux `ru_maxrss` after the isolated CSNA cell. One seed; not a 10-seed mean or a GPU measurement. Leave-one-in utility fitting is not included; a Cora one-plus-one probe unit (0.94 s support-only, 0.88 s candidate-added, 200 epochs) extrapolates $\mathcal{O}(3 + 3N_s)$ to ≈ 188 s on Cora (N_s=70) and ≈ 82 s on PubMed (N_s=30).

Dataset	K	Pool	Admission (s)	Student (s)	Total CSNA (s)	Admit-all (s)	Peak (MiB)
Cora	1	70	43.2	1.68	45.4	9.09	1091
	10	70	43.0	1.67	45.1	8.74	1092
PubMed	1	30	305	8.82	314	92.4	1808
	10	30	318	8.88	327	108	1846

The trade-off is selective admission in scarce-label regimes against a many-fold end-to-end cost. On this CPU audit, CSNA is not cheaper than admit-all, and peak memory is comparable (≈ 1.1 GiB on Cora, ≈ 1.8 GiB on PubMed). Student times remain similar because rejected candidates are dropped before the final GCN; the extra wall-clock is admission. These cells do not re-estimate headline accuracy and are not an equal-compute comparison.

For a binary target z_i and predictions r_i, the common audit metrics include

$$\widehat{\mathrm{Brier}} = \frac{1}{N}\sum_i (r_i - z_i)^2, \tag{10}$$

$$\widehat{\mathrm{ECE}} = \sum_{b:\,|I_b|>0} \frac{|I_b|}{N}\left|\frac{1}{|I_b|}\sum_{i\in I_b} z_i - \frac{1}{|I_b|}\sum_{i\in I_b} r_i\right|. \tag{11}$$

AUROC measures ranking; AUPRC depends on positive-target prevalence. ECE additionally depends on bins. The available aggregates do not identify all bins or candidate-level predictions, so no reliability diagram, calibration interval, or class-conditional audit is reconstructed from them.

D Complete Principal Metrics and Proxy Audit

Table 11: Principal macro-F1 (%, mean±std, 10 seeds). Differences are computed from the displayed means. Negative differences on ogbn-arxiv are retained and discussed in the main text.

Dataset	K	Admit-all	CSNA	Δ
Cora	1	68.62±3.05	76.12±2.06	+7.50
	2	69.84±5.44	74.92±2.37	+5.08
	5	74.82±1.30	77.67±1.60	+2.85
	10	77.43±2.20	78.90±1.89	+1.47
PubMed	1	68.90±5.23	74.19±3.62	+5.29
	2	71.41±4.74	75.43±3.31	+4.02
	5	74.94±2.70	76.68±2.22	+1.74
	10	77.48±2.23	77.63±1.82	+0.15
Citeseer	1	23.58±7.74	41.44±4.66	+17.86
	2	36.82±4.02	50.99±1.41	+14.17
	5	45.01±3.53	54.53±3.43	+9.52
	10	53.31±3.20	56.15±3.27	+2.84
ogbn-arxiv	1	37.06±2.21	35.78±2.81	-1.28
	2	38.76±2.01	37.58±2.79	-1.18
	5	39.29±1.69	39.97±2.71	+0.68
	10	41.21±1.58	42.05±2.80	+0.84

Table 12: Paired accuracy differences for the principal comparison (CSNA minus admit-all, pp; $n = 10$ matched seeds). Δ is the seed-level paired mean; the 95% CI is the Student-t interval for that mean; $d_z = \mathrm{mean}(\mathrm{diff})/\mathrm{sd}(\mathrm{diff})$; W/L counts seeds with positive versus negative differences; p is the two-sided Wilcoxon signed-rank p-value. No multiplicity correction and no significance stars. Citeseer uses prototype features. These quantities are computed from the same seeds as Table 1, not from unpaired mean±std columns.

Dataset	K	Δ	95% CI	d_z	W/L	p
Cora	1	+6.38	[4.17, 8.59]	2.06	10/0	0.002
	2	+4.42	[1.96, 6.88]	1.28	9/1	0.006
	5	+2.40	[1.50, 3.30]	1.90	10/0	0.002
	10	+1.22	[0.58, 1.86]	1.36	9/1	0.006
PubMed	1	+5.33	[2.41, 8.25]	1.31	9/1	0.006
	2	+4.09	[1.52, 6.66]	1.14	10/0	0.002
	5	+1.86	[0.12, 3.60]	0.76	9/1	0.037
	10	+0.26	[−0.96, 1.48]	0.15	7/3	0.625
Citeseer	1	+18.23	[14.46, 22.00]	3.46	10/0	0.002
	2	+15.54	[12.58, 18.50]	3.75	10/0	0.002
	5	+9.87	[6.08, 13.66]	1.86	10/0	0.002
	10	+3.85	[2.51, 5.19]	2.06	10/0	0.002
ogbn-arxiv	1	+2.46	[0.76, 4.16]	1.03	9/1	0.014
	2	+1.99	[0.36, 3.62]	0.87	8/2	0.041
	5	+2.44	[1.45, 3.43]	1.76	9/1	0.004
	10	+2.04	[0.86, 3.22]	1.24	8/2	0.010

Table 13: Paired macro-F1 differences for the same principal seeds as Tables 1 and 12 (pp). Δ is the seed-level paired mean, which may differ at the second decimal from displayed-mean subtraction in Table 11. Columns otherwise follow Table 12. All four ogbn-arxiv intervals include zero, as does PubMed $K = 10$.

Dataset	K	Δ	95% CI	d_z	W/L	p
Cora	1	+7.50	[5.01, 9.99]	2.16	10/0	0.002
	2	+5.07	[1.63, 8.52]	1.05	9/1	0.004
	5	+2.84	[1.88, 3.80]	2.12	10/0	0.002
	10	+1.47	[0.63, 2.31]	1.25	9/1	0.006
PubMed	1	+5.29	[2.41, 8.17]	1.31	9/1	0.006
	2	+4.02	[1.50, 6.54]	1.14	10/0	0.002
	5	+1.74	[0.00, 3.48]	0.72	9/1	0.037
	10	+0.15	[−1.19, 1.50]	0.08	6/4	0.695
Citeseer	1	+17.86	[11.78, 23.93]	2.10	10/0	0.002
	2	+14.17	[11.85, 16.49]	4.37	10/0	0.002
	5	+9.52	[6.20, 12.84]	2.05	10/0	0.002
	10	+2.84	[1.67, 4.01]	1.73	10/0	0.002
ogbn-arxiv	1	−1.28	[−3.82, 1.26]	−0.36	3/7	0.375
	2	−1.18	[−3.40, 1.04]	−0.38	2/8	0.105
	5	+0.69	[−0.47, 1.84]	0.42	7/3	0.375
	10	+0.84	[−0.67, 2.35]	0.40	7/3	0.275

Table 14: Separate proxy/corruption audit, means over seeds. $\eta = 0$ uses an embedding proxy; positive η uses label-flip status. The exact seed count, candidate-level scores, binning, and fit/evaluation independence are not identified in this table. ρ is Spearman correlation.

Dataset	K	η	AUROC	AUPRC	ECE	Brier	ρ
cora	1	0.0	0.979	0.983	0.112	0.068	0.830
cora	1	0.1	0.997	1.000	0.056	0.031	0.517
cora	1	0.2	0.998	1.000	0.054	0.021	0.691
cora	1	0.4	0.996	0.998	0.054	0.025	0.841
cora	2	0.0	0.976	0.980	0.100	0.071	0.825
cora	2	0.1	0.997	1.000	0.051	0.026	0.517
cora	2	0.2	0.998	1.000	0.052	0.019	0.690
cora	2	0.4	0.997	0.998	0.050	0.024	0.843
pubmed	1	0.0	0.990	0.991	0.121	0.060	0.849
pubmed	1	0.1	0.989	0.999	0.102	0.066	0.508
pubmed	1	0.2	0.968	0.992	0.131	0.087	0.649
pubmed	1	0.4	0.957	0.974	0.150	0.089	0.776
pubmed	2	0.0	0.985	0.987	0.142	0.060	0.841
pubmed	2	0.1	0.983	0.998	0.087	0.060	0.502
pubmed	2	0.2	0.983	0.996	0.110	0.070	0.670
pubmed	2	0.4	0.953	0.972	0.110	0.077	0.770

The principal accuracy means are in Table 1; Table 11 supplies the corresponding macro-F1 means. Paired differences, 95% intervals, Wilcoxon tests, and d_z for the same seeds are in Tables 12 and 13. The accuracy benefit on ogbn-arxiv is not a substitute for the unfavorable macro-F1 outcomes at its two smallest supports, whose F1 intervals include zero.

Table 15: ogbn-arxiv class-frequency tertiles (mean over 10 seeds). Low = rarest third of classes by full-graph count. Δ is CSNA minus admit-all (pp). Ten candidates per class by construction.

K	Tertile	Admit rate (CSNA)	ΔF1	ΔRec	Zero-admit
1	low	32.6%	−5.01	−12.36	39.2%
	mid	77.1%	−2.62	+0.74	2.9%
	high	82.8%	+2.00	+4.77	0.0%
2	low	32.6%	−5.64	−11.86	33.1%
	mid	78.8%	−1.01	+1.01	2.9%
	high	85.6%	+1.23	+3.86	0.0%
5	low	33.5%	−1.82	−8.43	33.1%
	mid	80.5%	−0.48	−0.29	1.4%
	high	87.5%	+1.29	+3.05	0.0%
10	low	28.0%	−0.02	−6.25	46.9%
	mid	80.3%	−0.87	+0.06	1.4%
	high	88.6%	+1.53	+2.70	0.0%

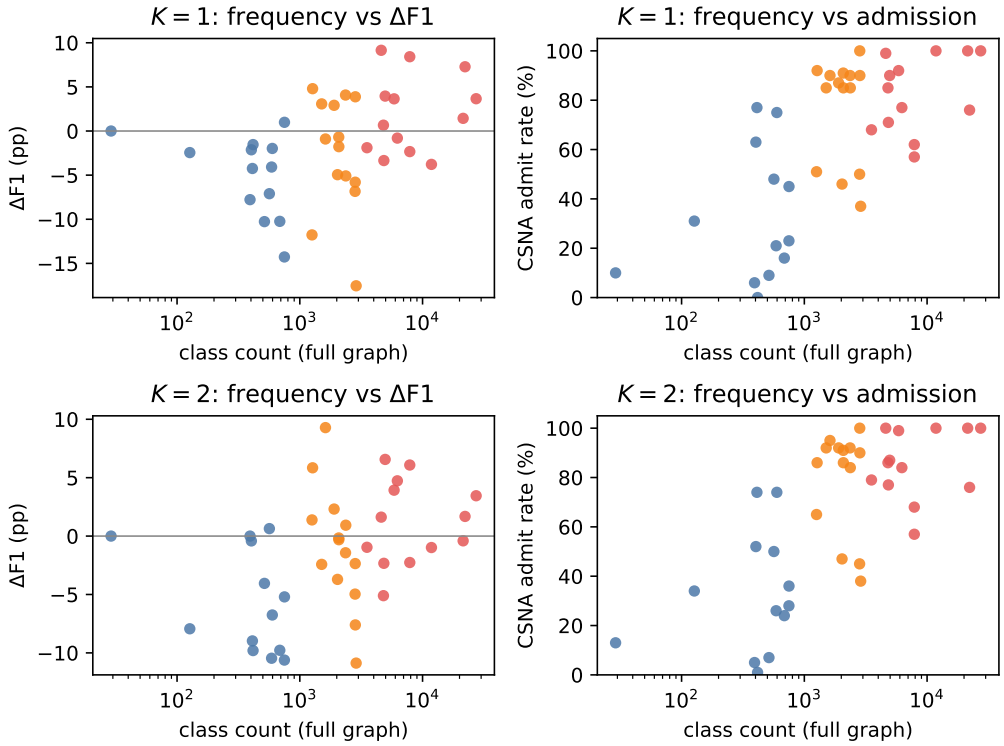

Figure 3: ogbn-arxiv class-wise diagnostic at $K = 1$ and $K = 2$: full-graph class count versus ΔF1 (left) and CSNA admission rate (right). Blue/orange/red mark rare/mid/frequent tertiles. Frequent classes receive higher admission and positive ΔF1; rare classes are under-admitted and lose F1.

Table 15 and Figure 3 come from a re-run of the paper-faithful protocol with extra per-class logging, not from reconstructing the frozen principal CSV. At $K = 1$, Spearman ρ between class frequency and CSNA admission rate is $+0.618$ (p=2.2×10^{-5}); versus ΔF1 it is $+0.427$ (p=0.006). Four classes are dropped in a majority of $K = 1$ seeds, and one class is never admitted. Class-balanced keep is not tested here as a replacement policy.

Table 16: Matched leave-one-out ablation of the principal protocol (accuracy %, mean±std, 10 seeds). Full clones `csna_hard_v21`. Utility-only is in Table 17.

Dataset	K	Full	−sem	−agree	−edge	−graph
Cora	1	78.65±1.65	78.51±1.36	79.04±1.58	78.42±2.10	78.79±1.89
	2	77.63±1.98	77.56±2.91	78.21±1.59	78.37±1.59	77.72±2.49
	5	79.48±1.71	80.01±2.33	79.66±2.30	79.60±1.97	80.12±1.57
	10	80.49±1.76	80.57±1.95	80.26±1.66	80.17±1.84	80.61±1.51
PubMed	1	74.39±3.73	73.29±4.42	72.51±4.16	75.07±3.26	73.49±4.01
	2	75.64±3.45	75.41±3.20	74.04±3.82	75.15±3.12	75.79±3.27
	5	76.86±2.36	77.32±1.73	74.30±3.76	77.53±1.95	76.76±2.07
	10	77.82±2.04	78.03±2.41	78.21±2.44	78.39±2.03	78.36±1.82
Citeseer	1	48.07±5.59	48.81±5.65	48.56±5.73	47.61±7.57	49.07±5.80
	2	56.67±1.64	55.93±2.91	55.80±3.71	56.67±1.85	56.28±3.59
	5	59.59±2.53	59.31±4.35	60.02±3.53	59.74±3.21	58.53±1.81
	10	61.15±3.92	61.34±3.72	61.51±3.68	61.11±2.33	60.95±3.74
ogbn-arxiv	1	54.49±3.87	55.30±3.55	55.82±2.27	54.22±3.97	55.25±3.14
	2	56.75±2.63	57.53±2.55	56.67±3.57	57.24±2.36	56.67±3.22
	5	58.04±2.26	58.17±3.11	58.75±2.25	57.23±2.58	57.93±2.79
	10	58.80±1.43	59.50±1.61	59.87±0.87	58.25±1.42	58.50±1.81

Table 17: Matched grouped-signal ablation of the same protocol as Table 16 (accuracy %, mean±std, 10 seeds). Semantic = (sem, agree); Structure = (edge, graph). Utility ranks on leave-one-in helpfulness only and is not a headline method. ogbn-arxiv skips utility (one GCN probe per synthetic).

Dataset	K	Full	Semantic	Structure	Utility
Cora	1	78.65±1.65	77.24±2.05	78.92±1.33	78.06±1.77
	2	77.63±1.98	77.10±2.33	78.37±1.88	78.39±1.93
	5	79.48±1.71	79.91±1.59	80.37±1.93	80.62±1.44
	10	80.49±1.76	80.15±1.60	80.56±2.13	81.14±1.49
PubMed	1	74.39±3.73	74.44±2.94	72.13±3.18	73.93±3.12
	2	75.64±3.45	76.23±3.15	73.88±4.11	74.99±3.92
	5	76.86±2.36	77.42±2.04	75.66±4.14	76.71±2.19
	10	77.82±2.04	78.39±2.57	77.89±2.47	78.13±1.63
Citeseer	1	48.07±5.59	47.65±6.58	49.25±3.95	49.44±5.07
	2	56.67±1.64	56.75±1.78	55.79±4.00	57.59±1.42
	5	59.59±2.53	58.58±2.18	60.05±3.31	61.16±2.15
	10	61.15±3.92	60.80±2.09	61.57±3.69	61.63±2.47
ogbn-arxiv	1	54.49±3.87	53.32±4.32	56.50±3.13	—
	2	56.75±2.63	56.22±2.79	57.23±3.28	—
	5	58.04±2.26	56.51±2.90	58.65±2.45	—
	10	58.80±1.43	57.50±1.31	59.46±0.91	—

Matched Full on Cora $K{=}1$ is 78.65 ± 1.65. PubMed without agreement is the largest Planetoid drop (-1.88 / -2.56 at $K{=}1/5$). Semantic-only is weaker than Full on Cora $K{=}1$ and on all four ogbn cells; structure-only is 56.50 versus Full 54.49 at ogbn $K{=}1$. Matched Full macro-F1 remains 35.77 / 37.52 at $K{=}1/2$ against admit-all 37.06 / 38.76.

Table 18: Equal validation-label budget (accuracy %, mean±std, 5 seeds). Both methods see the same n_{val} labels. CSNA splits $50/50$ (tune / stop); admit-all early-stops on the full mask.

Dataset	K	Method	10	25	50	100	250	500
Cora	1	Admit-all	57.86±15.37	67.92±4.58	67.86±4.87	71.96±1.61	72.30±2.21	72.66±2.00
		CSNA	42.66±14.42	71.50±7.70	74.62±4.70	76.96±2.57	79.50±1.14	79.60±1.04
	5	Admit-all	71.04±7.68	76.78±1.11	77.08±0.61	77.10±0.87	77.78±0.94	77.52±1.24
		CSNA	65.78±10.64	75.94±3.69	76.58±3.00	77.26±3.73	80.06±1.56	80.48±1.52
PubMed	1	Admit-all	62.88±6.99	65.28±3.22	65.22±3.59	65.18±4.07	66.72±2.61	66.44±2.32
		CSNA	57.66±7.34	61.88±12.25	69.46±3.03	71.66±2.46	74.40±2.59	74.46±2.24
	5	Admit-all	69.74±2.49	71.14±3.67	72.22±2.18	71.98±2.92	72.92±2.52	73.22±2.08
		CSNA	66.16±2.40	68.58±3.32	70.52±3.48	73.08±4.14	76.40±2.41	76.54±1.80

At $n_{\text{val}}{=}10$, CSNA is below admit-all in all four cells. At $n_{\text{val}}{=}25$ it is ahead only on Cora $K{=}1$ (71.50 versus 67.92) and behind on Cora $K{=}5$ and both PubMed cells. At $n_{\text{val}}{=}50$ it is ahead on both $K{=}1$ cells and behind on both $K{=}5$ cells. The $n_{\text{val}}{=}500$ Cora $K{=}1$ CSNA mean 79.60 ± 1.04 is a 5-seed cell of this sweep; it is not used to overwrite the 10-seed headline 78.65 ± 1.65.

Table 14 is a different audit from the tune-half utility diagnostic in Table 5. At $\eta = 0$, it uses an embedding-based target; at positive corruption rates, it uses known label-flip status. A node whose label was not flipped is not necessarily factually correct or useful. Without aligned score vectors, target definitions, and evaluation indices, the high embedding-proxy AUROC here and the low embedding-proxy AUROC in Table 5 do not establish one coherent independent utility-calibration experiment.

E COMPLETE ENCODER, GENERATOR, AND BACKBONE GRIDS

Table 19: Text-encoder study: test accuracy (%, mean±std; 10 seeds). Each encoder re-embeds the same raw text. The within-study means and differences are not pooled with the principal run.

Dataset	K	Encoder	Admit-all	CSNA	Δ
cora	1	bge-base	68.37±4.71	74.39±2.65	+6.02
cora	1	e5-base	68.71±3.00	76.73±2.46	+8.02
cora	1	minilm	71.87±2.70	79.29±1.41	+7.42
cora	1	mpnet	64.23±5.76	75.06±2.77	+10.83
cora	5	bge-base	76.20±1.77	78.28±2.27	+2.08
cora	5	e5-base	77.82±1.49	79.14±1.99	+1.32
cora	5	minilm	77.79±1.30	79.65±2.05	+1.86
cora	5	mpnet	76.74±2.06	79.08±1.56	+2.34
pubmed	1	bge-base	72.80±3.52	77.57±4.00	+4.77
pubmed	1	e5-base	76.04±4.56	79.76±1.95	+3.72
pubmed	1	minilm	68.93±5.13	74.60±3.95	+5.67
pubmed	1	mpnet	66.80±5.02	73.30±4.05	+6.50
pubmed	5	bge-base	77.59±2.91	79.56±3.34	+1.97
pubmed	5	e5-base	78.20±2.44	80.03±2.51	+1.83
pubmed	5	minilm	75.00±2.69	76.86±2.36	+1.86
pubmed	5	mpnet	72.31±3.46	75.58±3.78	+3.27

Table 20: Candidate-generator study: test accuracy (%, mean±std; 10 seeds). Admit-all uses the generator on each row; prototype and template candidates are not mislabeled as LLM-generated text.

Dataset	K	Generator	Admit-all	CSNA	Δ
cora	1	llm4ng	72.27±2.44	78.65±1.65	+6.38
cora	1	prototype	53.92±5.82	64.48±7.15	+10.56
cora	1	template	73.58±1.83	77.43±0.60	+3.85
cora	5	llm4ng	77.08±1.09	79.48±1.71	+2.40
cora	5	prototype	75.59±2.01	77.76±1.98	+2.17
cora	5	template	78.02±0.75	79.85±1.23	+1.83
pubmed	1	llm4ng	69.06±5.08	74.39±3.73	+5.33
pubmed	1	prototype	56.80±10.86	58.78±8.98	+1.98
pubmed	1	template	68.50±7.12	74.84±1.95	+6.34
pubmed	5	llm4ng	75.00±2.69	76.86±2.36	+1.86
pubmed	5	prototype	65.38±6.23	68.13±4.70	+2.75
pubmed	5	template	74.45±3.56	75.97±4.38	+1.52

Table 21: Full backbone study: mean accuracy differences (CSNA minus admit-all, pp; 5 seeds). The two negative outcomes are not omitted.

Dataset	K	APPNP	GAT	GCN	GraphSAGE
cora	1	+6.22	+6.74	+6.94	+13.72
cora	2	+4.58	+5.22	+4.44	+9.96
cora	5	+2.42	+2.48	+2.96	+5.18
cora	10	+1.46	+1.74	+1.26	+4.38
pubmed	1	+3.74	+8.56	+8.02	+8.42
pubmed	2	+1.90	+7.82	+6.52	+8.42
pubmed	5	+0.22	+5.06	+3.32	+5.80
pubmed	10	-2.38	-0.32	+0.40	+1.78

Each grid is evaluated with its own within-study comparator. Encoder shifts re-embed the same text; generator shifts replace the pool. The column "Admit-all" denotes that policy applied to the generator listed on the row, not LLM4NG-generated text for every row. Prototype pools are feature-level candidates. Interface portability does not establish that a single fitted score model transfers unchanged.

F Additional Comparative and Corruption Results

Table 22: Complete label-corruption study: test accuracy (%, 10-seed means). Each method receives the same flipped-label pool within a condition. All zero-noise values remain those of this sweep.

Dataset	K	η	Admit-all	CSNA	Δ
Cora	1	0.0	72.19	78.91	+6.72
Cora	1	0.1	69.78	79.68	+9.90
Cora	1	0.2	67.38	79.58	+12.20
Cora	1	0.4	61.88	79.68	+17.80
Cora	5	0.0	77.08	79.48	+2.40
Cora	5	0.1	76.82	80.12	+3.30
Cora	5	0.2	76.50	80.73	+4.23
Cora	5	0.4	73.62	81.19	+7.57
PubMed	1	0.0	69.06	74.39	+5.33
PubMed	1	0.1	68.05	74.95	+6.90
PubMed	1	0.2	67.03	73.80	+6.77
PubMed	1	0.4	62.89	72.33	+9.44
PubMed	5	0.0	75.00	76.86	+1.86
PubMed	5	0.1	74.57	77.36	+2.79
PubMed	5	0.2	72.78	76.81	+4.03
PubMed	5	0.4	69.15	74.61	+5.46

Table 23: Mixed realistic-failure pool at rate 0.4: test accuracy (%, 10-seed means). Corrupted synthetics mix plausible wrong-class labels, ambiguous features, forced wrong-class attachment, support duplicates, and feature noise. Clean rows are the principal reproducibility track, not Table 22. Descriptive gaps only; no Holm correction.

Dataset	K	Rate	Admit-all	CSNA	Δ
Cora	1	0.0	72.27	78.65	+6.38
Cora	1	0.4	70.53	77.74	+7.21
Cora	5	0.0	77.08	79.48	+2.40
Cora	5	0.4	76.30	80.05	+3.75
PubMed	1	0.0	69.06	74.39	+5.33
PubMed	1	0.4	68.37	73.24	+4.87
PubMed	5	0.0	75.00	76.86	+1.86
PubMed	5	0.4	73.75	77.59	+3.84

Table 24: Separate admission ladder: accuracy (%, mean±std; 10 seeds). Admit-all differs from the principal track; these comparisons are not merged. Proxy is an embedding-utility selector, not an oracle upper bound.

Dataset	K	No synthetic	Admit-all	Random	CSNA	Proxy
cora	1	57.76±8.36	78.29±1.97	77.90±2.08	78.65±1.65	73.56±2.00
cora	2	65.66±7.23	78.10±1.89	78.28±2.78	77.65±1.97	74.98±2.56
cora	5	76.51±2.12	80.54±1.15	80.17±1.50	79.48±1.71	78.78±0.85
cora	10	79.62±1.53	81.05±1.69	80.90±1.61	80.49±1.76	80.03±1.81
pubmed	1	55.84±10.09	71.29±3.58	70.94±3.99	74.39±3.73	64.38±7.65
pubmed	2	59.67±7.28	73.79±4.29	72.82±4.67	75.64±3.45	67.11±6.22
pubmed	5	66.24±5.70	77.24±2.03	75.62±2.61	76.86±2.36	69.87±6.38
pubmed	10	73.96±3.65	78.03±1.45	78.45±1.57	77.82±2.04	75.45±3.31

Table 25: Cora $K = 1$ exploratory probe. Accuracy (%); seed count and variability are not recorded. No significance or comparability with the principal absolute scores is inferred.

Method	Accuracy
MLP-only (support embeddings)	45.30
No-Synthetic GCN	48.00
Self-Training (LLM4NG pool)	80.20
CSNA (Ours)	80.70

The corruption table retains the zero-noise values from its own sweep. The robust-loss table is not merged with the separately reported baseline-suite contrasts in Table 2. In particular, repeated rounded CSNA and combination means do not identify a causal filtering mechanism.

The admission ladder is displayed in full, including conditions where admit-all or random retention is higher than CSNA. Its admit-all baselines differ from the principal comparison; this is not repaired by copying principal means into the ladder or subtracting a common offset. The historically named "oracle" ranks candidates using an embedding-space utility proxy and is not an upper bound. The extra probe has no recorded seed count or uncertainty and is descriptive only.

G Historical Archive: Unmatched Diagnostics and Reproduction Grid

The following tables are a historical archive. They are not part of the main pipeline and are not used to explain Table 1.

Table 26: Configuration-unmatched subset sweep: best displayed mean per cell (%, 10 seeds). This table is retained as a diagnostic, not used to explain the principal CSNA result or to select a policy on the test set.

Dataset	K	Best row within this sweep	Accuracy
Cora	1	CSNA-Graph	73.23 ± 2.05
Cora	2	CSNA-Graph	73.99 ± 3.35
Cora	5	CSNA-Graph	77.90 ± 1.13
Cora	10	CSNA-Semantic	79.70 ± 1.85
PubMed	1	CSNA-Agreement+Edge+Graph	72.96 ± 2.42
PubMed	2	CSNA-Edge+Graph	73.54 ± 3.46
PubMed	5	CSNA-Agreement	75.38 ± 2.09
PubMed	10	LLM4NG	78.70 ± 2.18

Table 27: Configuration-unmatched leave-one-out study: accuracy (%, means over 10 seeds). Full denotes the full variant in this study only. It is not substituted for the principal CSNA model.

Dataset	K	Full	$-$sem	$-$agree	$-$edge	$-$graph
cora	1	71.05	70.61	70.56	70.70	70.54
cora	2	71.99	71.48	71.61	71.81	71.47
cora	5	77.37	77.42	77.60	77.36	76.99
cora	10	79.58	79.36	79.24	79.44	79.70
pubmed	1	71.50	72.56	69.73	71.20	69.88
pubmed	2	71.64	73.30	70.87	70.93	71.22
pubmed	5	74.05	74.89	74.18	74.13	73.49
pubmed	10	77.53	78.03	77.89	77.52	77.57

Table 27's Cora $K{=}1$ Full is 71.05; the matched ablation used in the main text is 78.65. These are different studies.

Table 28: Separate ogbn-arxiv reproduction grid, low-shot panel (accuracy %, mean±std; 5 seeds). The $K = 0$ algorithm is not specified by the positive-shot method. DGI/MVGRL/GRACE/Meta-PN are lite implementations.

Method	$K = 0$	1	2	3
GCN (PyG)	—	23.44±3.87	29.56±4.00	34.32±3.16
GCN (MiniLM)	—	31.06±5.56	38.78±2.76	42.04±5.22
GAT (PyG)	—	26.98±3.08	35.30±3.00	38.80±4.33
GAT (MiniLM)	—	30.02±5.77	43.34±2.07	47.66±3.46
DGI	—	22.68±4.17	30.10±2.16	32.40±2.82
MVGRL	—	26.16±3.22	35.82±4.07	37.92±2.52
GRACE	—	27.20±3.47	38.76±2.91	39.22±3.80
Mixup+GCN	—	24.02±4.44	30.22±4.25	35.10±3.65
DropEdge+GCN	—	23.86±2.92	31.62±3.92	35.80±3.00
Meta-PN	—	8.08±4.32	8.86±4.50	42.30±4.97
CGPN	—	OOM	OOM	OOM
LLM4NG+GCN	50.40±2.04	50.76±3.47	54.10±2.45	54.18±3.44
LLM4NG+GAT	50.92±0.96	51.44±2.11	53.84±3.51	52.80±3.23
LLM4NG+GCN w/o A	44.36±3.11	48.46±3.14	51.66±2.18	50.40±4.36
LLM4NG+GAT w/o A	52.78±2.11	52.60±3.54	56.72±2.66	55.14±3.35
CSNA	52.70±1.81	53.46±3.40	56.38±3.01	56.16±2.24

Table 29: Separate ogbn-arxiv reproduction grid, remaining columns (accuracy %, mean±std; 5 seeds). w/o A denotes no synthetic attachment. These are within-grid results, not comparisons against the published LLM4NG scores.

Method	$K = 4$	5	10
GCN (PyG)	38.42±2.93	40.76±2.17	43.98±1.55
GCN (MiniLM)	45.98±3.90	49.20±3.52	53.76±2.70
GAT (PyG)	42.34±3.42	45.42±3.34	48.92±1.86
GAT (MiniLM)	48.22±2.43	51.62±2.62	53.64±1.10
DGI	36.30±1.15	39.56±1.79	42.10±0.99
MVGRL	41.26±2.76	43.22±3.41	50.38±1.89
GRACE	42.34±2.08	44.32±4.37	49.70±1.09
Mixup+GCN	38.00±2.78	40.82±3.28	43.80±1.11
DropEdge+GCN	38.10±2.39	41.96±1.09	45.42±1.96
Meta-PN	45.82±1.38	49.46±3.93	54.78±1.81
CGPN	OOM	OOM	OOM
LLM4NG+GCN	54.00±2.84	55.68±3.00	56.56±0.49
LLM4NG+GAT	53.92±1.55	53.92±1.77	54.42±1.25
LLM4NG+GCN w/o A	53.56±3.63	54.84±2.46	57.20±0.78
LLM4NG+GAT w/o A	54.78±2.12	57.04±3.05	56.24±1.80
CSNA	55.88±1.53	58.76±2.36	58.90±1.52

The separate ogbn-arxiv grid uses 5 seeds and includes $K = 0$. DGI, MVGRL, GRACE, and Meta-PN are scale-adapted lite implementations; CGPN is recorded as out of memory for positive-shot cells. Hardware and exact lite modifications are not specified, so OOM is a property of these runs rather than a universal limitation. The grid is not pooled with the 10-seed positive-shot principal track, and no superiority over the published LLM4NG results is asserted. CSNA is also not the highest row at every column. Its $K = 0$ result cannot be interpreted through the support-prototype method without the missing zero-shot implementation.

H Evidence Scope and Reporting Decisions

Reference track, not a preferred outcome. The principal four-dataset, positive-shot comparison is the reference for CSNA-versus-admit-all statements because it states the shared pool, features, splits, and attachment rule and contains both accuracy and macro-F1. Each other study keeps its own baseline. This choice does not invalidate an unfavorable result elsewhere: adverse selector, budget,

macro-F1, and backbone findings remain in the main text, and the separate ladder and ablations remain above. No run is relabeled as matched, independent, or correctly implemented merely because its numbers are plausible.

Arithmetic corrections are not new measurements. Observed means and standard deviations are unchanged. When both component means are displayed, differences in mean tables are computed from those displayed means, not invented full-precision values. PubMed $K = 10$ is therefore $77.82 - 77.56 = +0.26$, replacing the earlier prose value $+0.28$. The corresponding macro-F1 differences are $+5.08$ for Cora $K = 2$, $+2.85$ for Cora $K = 5$, and $+0.68$ for ogbn-arxiv $K = 5$; these replace $+5.07$, $+2.84$, and $+0.69$ in the earlier presentation. Tables 12 and 13 instead use the seed-level paired mean, so those three F1 cells appear as $+5.07$, $+2.84$, and $+0.69$ there. The portability means are arithmetic summaries of the displayed per-cell differences. A discrepancy potentially attributable to hidden precision is not declared an error in the underlying experiment; the reporting convention is simply made explicit.

Unmatched means. Appendix G keeps historical numbers, including unmatched Full 71.05 and the ladder admit-all 78.29 ± 1.97. They are not aligned with headline 78.65 / 72.27 and are not used to explain them.

Statistical claims. For the principal CSNA-versus-admit-all track, seed-level test accuracy and macro-F1 are available. Tables 12 and 13 report paired mean differences, 95% Student-t intervals, two-sided Wilcoxon tests, Cohen's d_z, and seed win/loss counts computed from those matched seeds. These quantities are not reconstructed from unpaired mean$\pm$std columns. No Holm or other multiplicity correction is applied, and no cell is starred as significant. Mixed-track Holm flags from other reports are not used. Standard deviations in mean tables remain variability across seeds, not intervals on a difference. An interval that includes zero is not evidence of equivalence. The retained baseline-suite contrasts are explicitly study-specific reported differences where component means are unavailable, and they are not accompanied by paired tests.

Conceptual corrections. Pruning is not down-weighting; temperature does not change a fixed top-fraction ranking; additional validation is not a CK-only budget. The paper claims selective admission in scarce or unreliable-label regimes, not general dominance and not certified risk control. Held-out utility (Table 6) is the ranking claim; in-sample 0.786 stays in Table 5.

Remaining artifact-level limits. Exact split/seed masks, pool sizes and candidate texts/features, checkpoints and preprocessing, attachment choices, normalization/tie handling, proxy-score fit/evaluation indices, ECE bins, and complete run configurations are required for independent reproduction. These are not inferred from aggregate tables. Accordingly, the manuscript presents an evidence-bounded empirical and analytical account rather than an independently replicated experiment package.